Health Intelligence Center

Division of Health Medical Data Science
健康医療データサイエンス分野

Professor	Seiya Imoto, Ph.D.	教　授　博士(数理学)	井　元　清　哉
Assistant Professor	Takanori Hasegawa, Ph.D.	助　教　博士(情報学)	長谷川　嵩　矩

Our mission is to utilize genomic big data and time series health medical data to realize methods for prediction and prevention of diseases and keeping/improving our health. For this purpose, we develop novel computational data analysis technologies by integrating Bayesian statistical theory and high-performance computing on supercomputer system.

1. Development of Computational Platform for Clinical Sequence and Interpretation

Shimizu E, Kasajima R, Yamaguchi K, Yokoyama K, Komura M, Saito A, Kobayashi M, Yuji K, Takane K, Shibuya T, Hasegawa T, Miyagi Y, Muto K, Tojo A, Furukawa Y, Miyano S, Yamaguchi R, Imoto S

From April 2015, Medical Genomics Research Initiative The University of Tokyo is launched. For implementing clinical sequence in the Institute of Medical Science, we formed a team of researchers and technicians who have various academic backgrounds including medicine, biology, pharmacology, genetics, statistics, computer science, ethics, etc. A highly secure infrastructure for analyzing personal genome was constructed; in the space, next generation sequencers are directly connected to a part (disconnect to internet) of supercomputer system in Human Genome Center and, for keeping traceability, laboratory information management system (LIMS) is installed to record all logs of wet experiments and computational analyses. Together with genome analysis in clinical sequence, we now intensively focus on a method for interpreting personal genome information. In July 2015, we started to use IBM Watson for cancer research to interpret the results of genome analyses. The results of genome sequence analysis including the interpretation of IBM Watson are evaluated and discussed in biweekly sequence board meeting. In 2016, we analyzed around sequence data of 100 cancer patients (more than 250 sequencing samples) with whole genome, exome, target deep sequencings. Also, multi-omics data including genome, transcriptome and epigenome were measured for integrative analysis that has the potential to achieve highly precise interpretation. This research is also performed as a part of the University of Tokyo's Center of Innovation (COI) project "Self-Managing Healthy Society".

2. Health Medical big data analysis

a. Integration of the records of health examination, microbiome and genomic data for predicting disease risks

Hasegawa T, Kakuta M, Yamaguchi R, Imoto S

Owing to increasing medical expenses, researchers have attempted to grasp clinical signs and preventive measures of diseases using electronic health record (EHR). In particular, time-series EHRs collected by periodic medical check-up enable us to

clarify the relevance among check-up results and individual environmental factors such as lifestyle. However, usually such time-series data have many missing observations and some results are strongly correlated to each other. These problems make the analysis difficult and there exists strong demand to detect clinical findings beyond them.

We focus on blood test values in medical check-up results and apply a time-series analysis methodology using a state space model. It can infer the internal medical states emerged in blood test values and handle missing observations. The estimated models enable us to predict one's blood test values under specified condition and predict the effect of intervention, such as changes of body composition and lifestyle.

We use time-series data of EHRs periodically collected in the Hirosaki cohort study in Japan and elucidate the effect of 17 environmental factors to 38 blood test values. Using the estimated model, we then simulate and compare time-transitions of participant's blood test values under several lifestyle scenarios. It visualizes the impact of lifestyle changes for the prevention of diseases. Finally, we exemplify that prediction errors under participant's actual lifestyle can be partially explained by genetic variations, and some of their effects have not been investigated by traditional association studies.

3. Computational Methods in Systems Biology and Immunology

a. Adaptive NetworkProfiler for identifying cancer characteristic-specific gene regulatory networks.

Park H[1], Shimamura T[2], Imoto S, Miyano S: [1]Faculty of Global and Science Studies, Yamaguchi University, [2]Graduate School of Medicine, Nagoya University

There is currently much discussion about sample (patient)-specific gene regulatory network identification, since the efficiently constructed sample-specific gene networks lead to effective personalized cancer therapy. Although statistical approaches have been proposed for inferring gene regulatory networks, the methods cannot reveal sample-specific characteristics because the existing methods, such as an L1-type regularization, provide averaged results for all samples. Thus, we cannot reveal sample-specific characteristics in transcriptional regulatory networks. To settle on this issue, the NetworkProfiler was proposed based on the kernel-based L1-type regularization. The NetworkProfiler imposes a weight on each sample based on the Gaussian kernel function for controlling effect of samples on modeling a target sample, where the amount of weight depends on similarity of cancer characteristics between samples. The method, however, cannot perform gene regulatory network identification well for a target sample in a sparse region (i.e., for a target sample, there are only a few samples having a similar characteristic of the target sample, where the characteristic is considered as a modulator in sample-specific gene network construction), since a constant bandwidth in the Gaussian kernel function cannot effectively group samples for modeling a target sample in sparse region. The cancer characteristics, such as an anti-cancer drug sensitivity, are usually nonuniformly distributed, and thus modeling for samples in a sparse region is also a crucial issue. We propose a novel kernel-based L1-type regularization method based on a modified k-nearest neighbor (KNN)-Gaussian kernel function, called an adaptive NetworkProfiler. By using the modified KNN-Gaussian kernel function, our method provides robust results against the distribution of modulators, and properly groups samples according to a cancer characteristic for sample-specific analysis. Furthermore, we propose a sample-specific generalized cross-validation for choosing the sample-specific tuning parameters in the kernel-based L1-type regularization method. Numerical studies demonstrate that the proposed adaptive NetworkProfiler effectively performs sample-specific gene network construction. We apply the proposed statistical strategy to the publicly available Sanger Genomic data analysis, and extract anti-cancer drug sensitivity-specific gene regulatory networks.

b. Bayesian model for analyzing human leukocyte antigen regions

Hayashi S, Yamaguchi R, Mizuno S[3], Komura M, Miyano S, Nakagawa H[4], Imoto S: [3]Center for Advanced Medical Innovation, Kyushu University, [4]RIKEN Center for Integrative Medical Sciences

Although human leukocyte antigen (HLA) genotyping based on amplicon, whole exome sequence (WES), and RNA sequence data has been achieved in recent years, accurate genotyping from whole genome sequence (WGS) data remains a challenge due to the low depth. Furthermore, there is no method to identify the sequences of unknown HLA types not registered in HLA databases. We developed a Bayesian model, called ALPHLARD, that collects reads potentially generated from HLA genes and accurately determines a pair of HLA types for each of HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1 genes at 3rd field resolution. Furthermore, ALPHLARD can detect rare germline variants not stored in HLA databases and call somatic mutations from paired normal and tumor sequence data. We illustrate the capability of ALPHLARD using 253 WES data and 25 WGS data

from Illumina platforms. By comparing the results of HLA genotyping from SBT and amplicon sequencing methods, ALPHLARD achieved 98.8% for WES data and 98.5% for WGS data at 2nd field resolution. We also detected three somatic point mutations and one case of loss of heterozygosity in the HLA genes from the WGS data. ALPHLARD showed good performance for HLA genotyping even from low-coverage data. It also has a potential to detect rare germline variants and somatic mutations in HLA genes. It would help to fill in the current gaps in HLA reference databases and unveil the immunological significance of somatic mutations identified in HLA genes.

c. An *in silico* automated pipeline to identify tumor specific neoantigens from next generation sequencing data

Hasegawa T, Hayashi S, Shimizu E, Mizuno S, Yamaguchi R, Miyano S, Nakagawa S, Imoto S:

Recent progress of massive parallel sequencing technology enables us to detect somatic mutations in each of cancer patients. It is known that some mutated peptides produced from missense mutations binds to the major histocompatibility complex (MHC). Since MHC presents mutated peptides to anti-tumor T cells, understanding this process is important in cancer immunotherapy. In this paper, we introduce a computational pipeline to predict binding affinity between mutated peptides and MHC molecules to detect neoantigens. We have implemented this pipeline on our supercomputer system. With nonsynonymous substitutions, frameshift insertions and deletions detected and intron retentions from whole-genome or exome sequencing data, we utilize RNA sequencing data and annotation data to make neoantigen detection pipeline more accurate.

4. Metagenome Analysis of Intestinal Microbiome

a. Analysis of intestinal microbiome.

Usui Y, Kimura Y, Satoh T, Takemura N, Ouchi Y, Ohmiya H, Kobayashi K, Suzuki H, Koyama S[5], Hagiwara S, Tanaka H, Imoto S, Eberl G[6], Asami Y[5], Fujimoto K, Uematsu S: [5]Food Science Research Laboratories, R&D Division, Meiji Co., Ltd, [6]Institut Pasteur, Microenvironment and Immunity Unit

The gut is an extremely complicated ecosystem where micro-organisms, nutrients and host cells interact vigorously. Although the function of the intestine and its barrier system weakens with age, some probiotics can potentially prevent age-related intestinal dysfunction. Lactobacillus delbrueckii subsp. bulgaricus 2038 and Streptococcus thermophilus 1131, which are the constituents of LB81 yogurt, are representative probiotics. However, it is unclear whether their long-term intake has a beneficial influence on systemic function. Here, we examined the gut microbiome, fecal metabolites and gene expression profiles of various organs in mice. Although age-related alterations were apparent in them, long-term LB81 yogurt intake led to an increased Bacteroidetes to Firmicutes ratio and elevated abundance of the bacterial family S24-7 (Bacteroidetes), which is known to be associated with butyrate and propanoate production. According to our fecal metabolite analysis to detect enrichment, long-term LB81 yogurt intake altered the intestinal metabolic pathways associated with propanoate and butanoate in the mice. Gene ontology analysis also revealed that long-term LB81 yogurt intake influenced many physiological functions related to the defense response. The profiles of various genes associated with antimicrobial peptides-, tight junctions-, adherens junctions- and mucus-associated intestinal barrier functions were also drastically altered in the LB81 yogurt-fed mice. Thus, long-term intake of LB81 yogurt has the potential to maintain systemic homeostasis, such as the gut barrier function, by controlling the intestinal microbiome and its metabolites.

Publications

1. Hayashi S, Moriyama T, Yamaguchi R, Mizuno S, Komura M, Miyano S, Nakagawa H, Imoto S. ALPHLARD-NT: Bayesian method for HLA genotyping and mutation calling through simultaneous analysis of normal and tumor whole-genome sequence data, *Journal of Computational Biology*, in press.
2. Hayashi S, Yamaguchi R, Mizuno S, Komura M, Miyano S, Nakagawa H, Imoto S. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data, *BMC Genomics*, 19(1): 790. doi: 10.1186/s12864-018-5169-9.
3. Muraoka D, Seo N, Hayashi T, Tahara Y, Fujii K, Tawara I, Miyahara Y, Okamori K, Yagita H, Imoto S, Yamaguchi R, Komura M, Miyano S, Goto M, Sawada S, Asai A, Ikeda H, Akiyoshi K, Harada N, Shiku H. Antigen delivery targeting tumor-associated macrophages overcomes tumor immune resistance, *Journal of Clinical Investigation*, in press.
4. Yokoyama K, Shimizu E, Yokoyama N, Naka-

mura S, Kasajima R, Ogawa M, Takei T, Ito M, Kobayashi A, Yamaguchi R, Imoto S, Miyano S, Tojo A. Cell-lineage level-targeted sequencing to identify acute myeloid leukemia with myelodysplasia-related changes, *Blood Advances*, DOI 10.1182/ bloodadvances.2017010744.

5. Nakamura S, Yokoyama K, Yusa N, Ogawa M, Takei T, Kobayashi A, Ito M, Shimizu E, Kasajima R, Wada Y, Yamaguchi R, Imoto S, Nagamura-Inoue T, Miyano S, Tojo A. Circulating tumor DNA dynamically predicts response and/ or relapse in patients with hematological malignancies. *Int J Hematol*. 108(4): 402-410, (2018). doi: 10.1007/s12185-018-2487-2.
6. Usui Y, Kimura Y, Satoh T, Takemura N, Ouchi Y, Ohmiya H, Kobayashi K, Suzuki H, Koyama S, Hagiwara S, Tanaka H, Imoto S, Eberl G, Asami Y, Fujimoto K, Uematsu S. Effects of long-term intake of a yogurt fermented with *Lactobacillus delbrueckii* subsp. bulgaricus 2038 and Streptococcus thermophilus 1131 on mice. *Int Immunol*. 30(7): 319-331, (2018). doi: 10.1093/ intimm/dxy035.
7. Inoue D, Fujino T, Sheridan P, Zhang Y-Z, Nagase R, Horikawa S, Li Z, Matsui H, Kanai A, Saika M, Yamaguchi R, Kozuka-Hata H, Kawabata K, Yokoyama A, Goyama S, Inaba T, Imoto S, Miyano S, Xu M, Yang F-C, Oyama M, and Kitamura T. A novel ASXL1-OGT axis plays roles in H3K4 methylation and tumor suppression in myeloid malignancies, *Leukemia*, 32, 1327-1337, (2018)
8. Park H, Shimamura T, Imoto S, Miyano S. Adaptive NetworkProfiler for identifying cancer characteristic-specific gene regulatory networks. *J Comput Biol*. 25(2): 130-145, (2018). doi: 10.1089/cmb.2017.0120.
9. VanderWeele DJ, Finney R, Katayama K, Gillard M, Paner G, Imoto S, Yamaguchi R, Wheeler D, Cam M, Maejima K, Sasaki-Oku A, Nakano K, Tanaka H, Pontier A, Grigoryev D, Kubo M, Ratain M, Miyano S, Nakagawa H. Local enrichment of multiple clinically significant genetic alterations within index foci of potentially lethal prostate cancer, *European Urology Focus*, (2018), pii: S2405-4569(18)30007-5. doi: 10.1016/j.euf.2018.01.006.
10. T. Hasegawa, K. Kojima, Y. Kawai, and M. Nagasaki. Time-series filtering for replicated observations via a kernel approximate Bayesian computation, *IEEE Transactions on Signal Processing*, 66(23), 6148-6161, (2018).
11. Fujita K, Chen X, Homma H, Tagawa K, Amano M, Saito A, Imoto S, Akatsu H, Hashizume Y, Kaibuchi K, Miyano S, Okazawa H. Targeting Tyro3 signals ameliorates PGRN-mutant FTLD-TDP model via tau-mediated synapse pathology, *Nature Communications*, 9: 433, (2018)
12. Ogawa M, Yokoyama K, Hirano M, Jimbo K, Ochi K, Kawamata T, Ohno N, Shimizu E, Yokoyama N, Yamaguchi R, Imoto S, Uchimaru K, Takahashi N, Miyano S, Imai Y, Tojo A. Different clonal dynamics of chronic myeloid leukaemia between bone marrow and the central nervous system, *British Journal of Haematology*, 183(5): 842-845 (2018). doi: 10.1111/bjh.15065
13. Kiyotani K, Mai T, Yamaguchi R, Yew P-Y, Kulis M, Orgel K, Imoto S, Miyano S, Burks A. W, Nakamura Y. Characterization of the B-cell receptor repertoires in peanut allergic subjects undergoing oral immunotherapy, *Journal of Human Genetics*, 63, 239-248, (2018)